

# Accurate distribution of $\mathbf{X}^T \mathbf{X}$ with singular, idempotent variance-covariance matrix $\mathbb{V}$

Hao Yuan Zhang and Jan Vrbik  
 Department of Mathematics  
 Brock University, 500 Glenridge Ave.  
 St. Catharines, Ontario, Canada, L2S 3A1

## Abstract

Assume that  $\mathbf{X}$  is a set of sample statistics which follow a special case Central Limit Theorem, namely: as the sample size  $n$  increases the corresponding distribution becomes multivariate Normal with the mean (of each  $X$ ) equal to zero and with an *idempotent* variance-covariance matrix  $\mathbb{V}$ . It is well known that  $\mathbf{X}^T \mathbf{X}$  has (in the same limit), a  $\chi^2$  distribution with degrees of freedom equal to the trace of  $\mathbb{V}$ . In this article we extend the above result to include the corresponding  $\frac{1}{n}$ -proportional corrections, making the new approximation substantially more accurate and extending its range of applicability to small-size samples.

## 1 Introduction

Consider a random independent sample of size  $n$  from a distribution, resulting in  $p$  sample statistics, say  $X_1, X_2, \dots, X_p$  (collectively denoted  $\mathbf{X}$ ). We assume that these are defined in such a way (visualize each being a function of standardized sample means) that their joint distribution converges (as  $n \rightarrow \infty$ ) to multivariate Normal. This implies that their joint  $j^{th}$ -order cumulants can be expanded in inverse powers of  $n$  as follows:

$$\mathcal{K}^{(j)} = \left( \kappa^{(j,0)} + \frac{\kappa^{(j,1)}}{n} + \frac{\kappa^{(j,2)}}{n^2} + \dots \right) n^{1-j/2}$$

where  $\mathcal{K}^{(j)}$  and each of the corresponding  $\kappa^{(j,\ell)}$  is a fully symmetric *tensor* with  $j$  (implicit) indices.

When  $j = 1$ ,  $\mathcal{K}^{(1)}$  is a (column) *vector* of the expected values of  $\mathbf{X}$ ; the  $\kappa^{(1,0)}$  term must always (rather exceptionally) equal to zero, and the  $\kappa^{(j,1)}$  term we rename  $\boldsymbol{\mu}$ , so that

$$\mathcal{K}^{(1)} = \frac{\boldsymbol{\mu}}{\sqrt{n}} + \dots$$

Similarly,  $\mathcal{K}^{(2)}$  is the corresponding variance-covariance *matrix*, expanded (and notationally simplified) as follows:

$$\mathcal{K}^{(2)} = \mathbb{V}^{(0)} + \frac{\mathbb{V}^{(1)}}{n} + \dots \quad (1)$$

We will now assume that  $\mathbb{V}^{(0)}$  is *idempotent* with a *trace* equal to  $k$ . Our task is to find the approximate distribution of

$$T = \mathbf{X}^T \mathbf{X} \quad (2)$$

to the  $\frac{1}{n}$  accuracy, extending the familiar result which states that, in the  $n \rightarrow \infty$  limit, the distribution of  $T$  becomes  $\chi^2$  with  $k$  degrees of freedom.

## 2 MGF and PDF of rotated $\mathbf{X}$

To find a more accurate approximation for  $T$ , we recall that there is an orthonormal matrix  $\mathbb{R}$  which diagonalizes  $\mathbb{V}^{(0)}$  thus:

$$\mathbb{R} \mathbb{V}^{(0)} \mathbb{R}^T \equiv \mathbb{H}$$

where  $\mathbb{H}$  is *main-diagonal*, with the first  $k$  diagonal elements equal to 1, the rest of them equal to 0; (2) can then be rewritten as

$$T = \mathbf{Z}^T \mathbf{Z}$$

where

$$\mathbf{Z} \equiv \mathbb{R} \mathbf{X}$$

Note that transforming a  $j^{\text{th}}$ -order cumulant of  $\mathbf{X}$  into its  $\mathbf{Z}$  counterpart is achieved by

$$\tilde{\mathcal{K}}_{i_1, i_2 \dots i_j} = \sum_{\ell_1, \ell_2 \dots \ell_j=1}^p \mathbb{R}_{i_1, \ell_1} \mathbb{R}_{i_2, \ell_2} \dots \mathbb{R}_{i_j, \ell_j} \mathcal{K}_{\ell_1, \ell_2 \dots \ell_j}$$

Also note that the  $\kappa^{(j, m)}$  (individually) transform in the same manner, and (essential to the rest of this article) that any component of the new  $\tilde{\kappa}^{(j, 0)}$  with a lower (*implicit*, in this notation) index in the  $k+1$  to  $p$  range must be equal to zero (no longer true for  $\tilde{\kappa}^{(j, 1)}$ ,  $\tilde{\kappa}^{(j, 2)}$  ..., including  $\tilde{\boldsymbol{\mu}}$  and  $\tilde{\mathbb{V}}^{(1)}$ ).

This implies that, in the  $n \rightarrow \infty$  limit, the  $Z_i$ s with  $i \leq k$  are independent, standardized Normal, and the remaining  $Z_i$ s ( $p-k$  of them) are identically equal to zero (implying the  $T \epsilon \chi_k^2$  result). But again, this happens only in the  $n \rightarrow \infty$  limit; both  $\frac{\tilde{\mu}_i}{\sqrt{n}}$  and the corresponding components of  $\frac{\tilde{\mathbb{V}}^{(1)}}{n}$  may remain *non-zero* even when  $i > k$  (similar to what happens to  $\frac{g(\bar{X}) - g(\mu)}{\sigma |g'(\mu)|}$  in the same limit, where  $\bar{X}$  is a sample mean and  $g$  is an arbitrary function). The  $\chi_k^2$  approximation thus becomes less accurate with decreasing  $n$ . To improve its accuracy, we proceed to find a  $\frac{1}{n}$ -proportional correction to it.

## 2.1 Expanding MGF

The corresponding cumulant generating function of  $\mathbf{Z}$ , expanded to the  $\frac{1}{n}$  accuracy, is given by

$$K(t_1, t_2, \dots, t_p) \equiv K(\mathbf{t}) = \frac{\mathbf{t}^T \mathbb{H} \mathbf{t}}{2} + \frac{\mathbf{t}^T \tilde{\boldsymbol{\mu}}^{(1)}}{\sqrt{n}} + \frac{\tilde{\kappa}^{(3,0)} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t}}{6\sqrt{n}} + \frac{\mathbf{t}^T \tilde{\mathbb{V}}^{(1)} \mathbf{t}}{2n} + \frac{\tilde{\kappa}^{(4,0)} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t}}{24n} + \dots$$

where the last ellipsis indicates terms beyond the  $\frac{1}{n}$  accuracy, and  $\tilde{\kappa} \circ \mathbf{t}$  implies contracting the last (implicit) index of  $\tilde{\kappa}$  with the only index of  $\mathbf{t}$ , i.e., explicitly,

$$\sum_{i_\ell=1}^p \tilde{\kappa}_{i_1, i_2, \dots, i_\ell} \cdot t_{i_\ell}$$

Thus, for example,  $\tilde{\kappa}^{(3,0)} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t}$  means that all 3 indices of  $\tilde{\kappa}^{(3,0)}$  have been contracted, one by one, with an index of  $\mathbf{t}$ , resulting in a scalar.

It is relatively easy to convert the above cumulant generating function into the corresponding moment generating function, thus

$$M(\mathbf{t}) = \exp[K(\mathbf{t})] = \exp\left(\frac{\sum_{i=1}^k t_i^2}{2}\right) \cdot \left(1 + \frac{\mathbf{t}^T \tilde{\mathbb{V}}^{(1)} \mathbf{t}}{2n} + \frac{\tilde{\kappa}^{(4,0)} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t}}{24n} + \frac{\left(\mathbf{t}^T \tilde{\boldsymbol{\mu}}^{(1)} + \frac{1}{6} \tilde{\kappa}^{(3,0)} \circ \mathbf{t} \circ \mathbf{t} \circ \mathbf{t}\right)^2}{2n} + \dots\right) \quad (3)$$

where this time we have discarded not only the  $o(\frac{1}{n})$  terms, but also terms with odd powers of  $\mathbf{t}$ ; they do not contribute to our final answer, as we will show shortly.

## 2.2 Corresponding PDF

The moment generating function (3) converts to the following joint PDF of the  $\mathbf{Z}$  distribution:

$$f(z_1, z_2, \dots, z_p) = (2\pi)^{-k/2} \cdot \left(1 + \frac{\mathbf{D}^T \tilde{\mathbb{V}}^{(1)} \mathbf{D}}{2n} + \frac{\tilde{\kappa}^{(4,0)} \circ \mathbf{D} \circ \mathbf{D} \circ \mathbf{D} \circ \mathbf{D}}{24n} + \frac{\left(\mathbf{D}^T \tilde{\boldsymbol{\mu}}^{(1)} + \frac{1}{6} \tilde{\kappa}^{(3,0)} \circ \mathbf{D} \circ \mathbf{D} \circ \mathbf{D}\right)^2}{2n} + \dots\right) \exp\left(-\frac{\sum_{i=1}^k z_i^2}{2}\right) \prod_{i=k+1}^p \delta(z_i) \quad (4)$$

where  $\delta(z_i)$  stands for the Dirac delta function (visualize it as a PDF of a Normal distribution with  $\mu = 0$  and  $\sigma \rightarrow 0$ ), and  $\mathbf{D}$  is a differential operator whose individual components are partial derivatives with respect to  $z_i$  (made more explicit later on).

### 3 Finding CDF of $T$

Let us now compute

$$\Pr(T < u) = \Pr\left(\sum_{i=1}^p Z_i^2 < u\right) \quad (5)$$

This can be done by integrating (4) over a sphere of radius  $\sqrt{u}$  in the  $p$ -dimensional space of the  $z_i$  variables. The main contribution is of course from

$$\begin{aligned} & (2\pi)^{-k/2} \int \cdots \int_{\mathcal{R}_p(\sqrt{u})} \exp\left(-\frac{\sum_{i=1}^k z_i^2}{2}\right) \prod_{i=k+1}^p \delta(z_i) dz_1 dz_2 \dots dz_p \\ &= (2\pi)^{-k/2} \int \cdots \int_{\mathcal{R}_k(\sqrt{u})} \exp\left(-\frac{\sum_{i=1}^k z_i^2}{2}\right) dz_1 dz_2 \dots dz_k \end{aligned} \quad (6)$$

where  $\mathcal{R}_d(\rho)$  denotes a  $d$ -dimensional sphere of radius  $\rho$ . Introducing  $r \equiv \sqrt{\sum_{i=1}^k z_i^2}$  and utilizing Fisher's geometrical method of multidimensional integration, (6) equals to

$$\begin{aligned} F(u) &\equiv (2\pi)^{-k/2} \int_0^{\sqrt{u}} S_k(r) \exp\left(-\frac{r^2}{2}\right) dr \\ &= (2\pi)^{-k/2} \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)} \int_0^{\sqrt{u}} r^{k-1} \exp\left(-\frac{r^2}{2}\right) dr \\ &= \frac{1}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \int_0^u w^{k/2-1} \exp\left(-\frac{w}{2}\right) dw \end{aligned} \quad (7)$$

where

$$S_k(r) = \frac{2\pi^{k/2} r^{k-1}}{\Gamma\left(\frac{k}{2}\right)}$$

is the area of the surface of a  $k$ -dimensional sphere of radius  $r$ , and (7) is clearly the CDF of the  $\chi_k^2$  distribution.

#### 3.1 $\tilde{\mu}^{(1)}$ and $\tilde{\mathbb{V}}^{(1)}$ corrections

From what follows it is easy to see why terms of (4) having an odd power of  $D_i$  (for any one or more  $i$ ) must yield zero contribution to (5) - that goes for those we have already deleted, as well for most of those which are still explicitly a part of (4), such as, for example

$$\dots + \frac{D_1 \tilde{\mathbb{V}}_{12}^{(1)} D_2}{n} + \dots$$

etc. The only terms which remain are those containing  $D_i^2$ ,  $D_i^4$ ,  $D_i^6$ ,  $D_i^2 D_j^2$ ,  $D_i^4 D_j^2$ , and  $D_i^2 D_j^2 D_\ell^2$ , where  $i \neq j \neq k$ . Let us establish their contribution

to (5). Starting with  $D_i^2$ , we get (using  $D_1^2$  as a proxy; due to the obvious symmetry, all  $D_i^2$ , where  $1 \leq i \leq k$ , will contribute equally):

$$\begin{aligned}
& (2\pi)^{-k/2} \int \cdots \int_{\mathcal{R}_p(\sqrt{u})} \frac{\partial^2}{\partial z_1^2} \exp\left(-\frac{\sum_{i=1}^k z_i^2}{2}\right) \prod_{i=k+1}^p \delta(z_i) dz_1 dz_2 \dots dz_p \\
&= (2\pi)^{-k/2} \int \cdots \int_{\mathcal{R}_k(\sqrt{u})} \frac{\partial^2}{\partial z_1^2} \exp\left(-\frac{\sum_{i=1}^k z_i^2}{2}\right) dz_1 dz_2 \dots dz_k \\
&= (2\pi)^{-k/2} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{\partial^2 \exp(-\frac{z_1^2}{2})}{\partial z_1^2} \int \cdots \int_{\mathcal{R}_{k-1}(\sqrt{u-z_1^2})} \exp\left(-\frac{\sum_{i=2}^k z_i^2}{2}\right) dz_2 \dots dz_k dz_1 \\
&= (2\pi)^{-k/2} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{\partial^2 \exp(-\frac{z_1^2}{2})}{\partial z_1^2} \int_0^{\sqrt{u-z_1^2}} S_{k-1}(r) \cdot \exp(-\frac{r^2}{2}) dr dz_1 \\
&= -\frac{2\pi^{-1/2}}{2^{k/2}\Gamma(\frac{k-1}{2})} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{\partial \exp(-\frac{z_1^2}{2})}{\partial z_1} \cdot \frac{\partial}{\partial z_1} \int_0^{\sqrt{u-z_1^2}} r^{k-2} \exp(-\frac{r^2}{2}) dr dz_1 \\
&= -\frac{2\pi^{-1/2}}{2^{k/2}\Gamma(\frac{k-1}{2})} \exp(-\frac{u}{2}) \int_{-\sqrt{u}}^{\sqrt{u}} z_1^2 (u - z_1^2)^{(k-3)/2} dz_1 \\
&= -\frac{2}{k \cdot 2^{k/2}\Gamma(\frac{k}{2})} \cdot u^{k/2} \exp(-\frac{u}{2}) = -\frac{2u}{k} \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2}\Gamma(\frac{k}{2})} \tag{8}
\end{aligned}$$

To follow the derivation, one must be able to: (i) handle Dirac's delta function inside an integral, (ii) establish the intersection of a plane with a sphere in  $k$  dimensions, (iii) perform by-part integration, (iv) differentiate with respect to a parameter which appears in the upper limit of an integral, and (v) be familiar with integrals relating to Beta function.

Careful examination of (4) reveals that (8) needs to be added to (7) after being multiplied by

$$a \equiv \frac{\sum_{i=1}^k \left( \tilde{\mathbf{V}}_{i,i}^{(1)} + \tilde{\mu}_i^{(1)} \cdot \tilde{\mu}_i^{(1)} \right)}{2n} \tag{9}$$

As for the  $D_i^2$  contribution to (5) when  $i > k$  (note that for these values of  $i$  there is no contribution from the  $\tilde{\kappa}^{(3,0)}$  and  $\tilde{\kappa}^{(4,0)}$  terms): it is fairly obvious that adding the sum of squares of these  $p - k$  random variables with zero variance and the mean of  $\frac{\mu^{(1)}}{\sqrt{n}}$  will simply increase the value of  $T$  by

$$\frac{\sum_{i=k+1}^p \tilde{\mu}_i^{(1)} \cdot \tilde{\mu}_i^{(1)}}{n}$$

Since the contribution of  $\tilde{V}_{i,i}^{(1)}$  is the same as that of  $\mu_i^{(1)} \cdot \mu_i^{(1)}$  at the MGF level, it must be the same (to this level of approximation) at the PDF level; this implies that the only effect of the  $Z_i$  ( $i > k$ ) variables will be adding

$$d \equiv \frac{\sum_{i=k+1}^p \left( \tilde{V}_{i,i}^{(1)} + \tilde{\mu}_i^{(1)} \cdot \tilde{\mu}_i^{(1)} \right)}{n} \quad (10)$$

to  $T$ .

### 3.2 Corrections due to 3<sup>rd</sup> and 4<sup>th</sup>-order cumulants

To find the contribution of the  $D_1^4$  and  $D_1^6$  terms, we repeat the steps leading to (8), increasing the power of the  $\frac{\partial}{\partial z_1}$  derivative (to 4 and 6, respectively). This means that

$$\frac{\partial \exp(-\frac{z_1^2}{2})}{\partial z_1} = -z_1 \exp(-\frac{z_1^2}{2})$$

of the third last line of (8) needs to be replaced by

$$\frac{\partial^3 \exp(-\frac{z_1^2}{2})}{\partial z_1^3} = -(z_1^3 - 3z_1) \exp(-\frac{z_1^2}{2})$$

and by

$$\frac{\partial^5 \exp(-\frac{z_1^2}{2})}{\partial z_1^5} = -(z_1^5 - 10z_1^3 + 15z_1) \exp(-\frac{z_1^2}{2})$$

respectively, correspondingly modifying the rest.

For the  $D_1^4$  contribution we now get, in place of (8):

$$3 \left( \frac{2u}{k} - \frac{2u^2}{k(k+2)} \right) \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2} \Gamma(\frac{k}{2})} \quad (11)$$

According to (4), this needs to be multiplied by

$$\frac{\sum_{i=1}^k \left( \tilde{\kappa}_{i,i,i,i}^{(4,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{i,i,i}^{(3,0)} \right)}{24n} \quad (12)$$

before being added to (7).

Similarly,  $D_1^6$  leads to

$$15 \left( -\frac{2u}{k} + \frac{4u^2}{k(k+2)} - \frac{2u^3}{k(k+2)(k+4)} \right) \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2} \Gamma(\frac{k}{2})}$$

and is to be multiplied by

$$\frac{\sum_{i=1}^k \left( \kappa_{i,i,i}^{(3,0)} \right)^2}{72n} \quad (13)$$

To deal with  $D_1^2 D_2^2$  (which covers the case of any  $D_i^2 D_j^2$  with  $i \neq j$ ) we follow a similar procedure (skipping the first rather obvious step):

$$\begin{aligned}
& (2\pi)^{-k/2} \int \cdots \int_{\mathcal{R}_k(\sqrt{u})} \frac{\partial^2}{\partial z_1^2} \frac{\partial^2}{\partial z_2^2} \exp\left(-\frac{\sum_{i=1}^k z_i^2}{2}\right) dz_1 dz_2 \dots dz_k \\
&= (2\pi)^{-k/2} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{\partial^2 \exp(-\frac{z_1^2}{2})}{\partial z_1^2} \int_{-\sqrt{u-z_1^2}}^{\sqrt{u-z_1^2}} \frac{\partial^2 \exp(-\frac{z_2^2}{2})}{\partial z_2^2} \cdot \\
&\quad \int \cdots \int_{\mathcal{R}_{k-2}(\sqrt{u-z_1^2-z_2^2})} \exp\left(-\frac{\sum_{i=3}^k z_i^2}{2}\right) dz_3 \dots dz_k dz_2 dz_1 \\
&= (2\pi)^{-k/2} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{\partial^2 \exp(-\frac{z_1^2}{2})}{\partial z_1^2} \int_{-\sqrt{u-z_1^2}}^{\sqrt{u-z_1^2}} \frac{\partial^2 \exp(-\frac{z_2^2}{2})}{\partial z_2^2} \cdot \\
&\quad \int_0^{\sqrt{u-z_1^2-z_2^2}} S_{k-2}(r) \exp\left(-\frac{r^2}{2}\right) dr dz_2 dz_1 \\
&= -\frac{2\pi^{-1}}{2^{k/2}\Gamma(\frac{k-2}{2})} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{\partial^2 \exp(-\frac{z_1^2}{2})}{\partial z_1^2} \int_{-\sqrt{u-z_1^2}}^{\sqrt{u-z_1^2}} \frac{\partial \exp(-\frac{z_2^2}{2})}{\partial z_2} \cdot \\
&\quad \frac{\partial}{\partial z_2} \int_0^{\sqrt{u-z_1^2-z_2^2}} r^{k-3} \exp\left(-\frac{r^2}{2}\right) dr dz_2 dz_1 \\
&= -\frac{2\pi^{-1}}{2^{k/2}\Gamma(\frac{k-2}{2})} \cdot \exp\left(-\frac{u}{2}\right) \int_{-\sqrt{u}}^{\sqrt{u}} (z_1^2 - 1) \int_{-\sqrt{u-z_1^2}}^{\sqrt{u-z_1^2}} z_2^2 (u - z_1^2 - z_2^2)^{(k-4)/2} dz_2 dz_1 \\
&= -\frac{\pi^{-1/2}}{2^{k/2}\Gamma(\frac{k+1}{2})} \cdot \exp\left(-\frac{u}{2}\right) \int_{-\sqrt{u}}^{\sqrt{u}} (z_1^2 - 1)(1 - z_1^2)^{(k-1)/2} dz_1 \\
&= \frac{2(1 - \frac{u}{k+2})}{k \cdot 2^{k/2}\Gamma(\frac{k}{2})} \cdot u^{k/2} \exp\left(-\frac{u}{2}\right) = \left(\frac{2u}{k} - \frac{2u^2}{k(k+2)}\right) \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2}\Gamma(\frac{k}{2})}
\end{aligned}$$

This is yet to be multiplied - see (4) - by

$$\frac{\sum_{i \neq j}^k \left( \tilde{\kappa}_{i,i,j,j}^{(4,0)} + \tilde{\kappa}_{i,j,i,j}^{(4,0)} + \tilde{\kappa}_{i,j,j,i}^{(4,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{i,j,j}^{(3,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{j,i,j}^{(3,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{j,j,i}^{(3,0)} \right)}{24n}$$

which, *together* with (12) *multiplied* by 3 (how very convenient of (11) to provide

this extra factor) can be combined into

$$\frac{\sum_{i,j=1}^k \left( \tilde{\kappa}_{i,i,j,j}^{(4,0)} + \tilde{\kappa}_{i,j,i,j}^{(4,0)} + \tilde{\kappa}_{i,j,j,i}^{(4,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{i,j,j}^{(3,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{j,i,j}^{(3,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{j,j,i}^{(3,0)} \right)}{24n} \quad (14)$$

Due to the total symmetry of cumulants, (14) can be simplified further to

$$\frac{\sum_{i,j=1}^k \left( \tilde{\kappa}_{i,i,j,j}^{(4,0)} + 4\tilde{\mu}_i^{(1)} \tilde{\kappa}_{i,j,j}^{(3,0)} \right)}{8n} \quad (15)$$

In (15), we can now change the upper limit of the summation from  $k$  to  $p$ , as all those extra terms are equal to zero (as explained earlier). And, since the resulting expression is rotationally invariant, we can express it in terms of the old  $X$  cumulants thus

$$b \equiv \frac{\sum_{i,j=1}^p \left( \kappa_{i,i,j,j}^{(4,0)} + 4\mu_i^{(1)} \kappa_{i,j,j}^{(3,0)} \right)}{8n} \quad (16)$$

Dealing with  $D_1^4 D_2^2$  is now quite easy - one has to replace (in the previous derivation)

$$\frac{\partial^2 \exp(-\frac{z_1^2}{2})}{\partial z_1^2} = (z_1^2 - 1) \exp(-\frac{z_1^2}{2})$$

by

$$\frac{\partial^4 \exp(-\frac{z_1^2}{2})}{\partial z_1^4} = (z_1^4 - 6z_1^2 + 3) \exp(-\frac{z_1^2}{2})$$

and work out the details, getting

$$3 \cdot \left( -\frac{2u}{k} + \frac{4u^2}{k(k+2)} - \frac{2u^3}{k(k+2)(k+4)} \right) \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2} \Gamma(\frac{k}{2})}$$

yet to be multiplied by

$$\frac{\sum_{i \neq j}^k \left( \begin{aligned} & 2\tilde{\kappa}_{i,i,j}^{(3,0)} \tilde{\kappa}_{j,j,j}^{(3,0)} + 2\tilde{\kappa}_{i,j,i}^{(3,0)} \tilde{\kappa}_{j,j,j}^{(3,0)} + 2\tilde{\kappa}_{j,i,i}^{(3,0)} \tilde{\kappa}_{j,j,j}^{(3,0)} + \\ & \left( \tilde{\kappa}_{i,i,j}^{(3,0)} \right)^2 + \left( \tilde{\kappa}_{j,i,i}^{(3,0)} \right)^2 + \left( \tilde{\kappa}_{j,j,i}^{(3,0)} \right)^2 + 2\tilde{\kappa}_{i,j,j}^{(3,0)} \tilde{\kappa}_{j,i,j}^{(3,0)} + 2\tilde{\kappa}_{i,j,j}^{(3,0)} \tilde{\kappa}_{j,j,i}^{(3,0)} + 2\tilde{\kappa}_{j,i,j}^{(3,0)} \tilde{\kappa}_{j,j,i}^{(3,0)} \end{aligned} \right)}{72n} \quad (17)$$

We would hope that by now the reader can supply the details of the  $D_1^2 D_2^2 D_3^2$  derivation, getting the (at this point expected) result of

$$\left( -\frac{2u}{k} + \frac{4u^2}{k(k+2)} - \frac{2u^3}{k(k+2)(k+4)} \right) \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2} \Gamma(\frac{k}{2})}$$



Referring back to (4), the last expression has to be multiplied by

$$\frac{\sum_{i \neq j \neq \ell}^k \left( \begin{array}{c} \tilde{\kappa}_{i,j,\ell}^{(3,0)} \cdot (\tilde{\kappa}_{i,j,\ell}^{(3,0)} + \tilde{\kappa}_{i,\ell,j}^{(3,0)} + \tilde{\kappa}_{j,i,\ell}^{(3,0)} + \tilde{\kappa}_{j,\ell,i}^{(3,0)} + \tilde{\kappa}_{\ell,i,j}^{(3,0)} + \tilde{\kappa}_{\ell,j,i}^{(3,0)}) + \\ \tilde{\kappa}_{i,j,j}^{(3,0)} \cdot \tilde{\kappa}_{i,\ell,\ell}^{(3,0)} + \tilde{\kappa}_{i,j,j}^{(3,0)} \cdot \tilde{\kappa}_{\ell,i,\ell}^{(3,0)} + \tilde{\kappa}_{i,j,j}^{(3,0)} \cdot \tilde{\kappa}_{\ell,\ell,i}^{(3,0)} + \tilde{\kappa}_{j,i,j}^{(3,0)} \cdot \tilde{\kappa}_{i,\ell,\ell}^{(3,0)} + \\ \tilde{\kappa}_{j,i,j}^{(3,0)} \cdot \tilde{\kappa}_{\ell,i,\ell}^{(3,0)} + \tilde{\kappa}_{j,i,j}^{(3,0)} \cdot \tilde{\kappa}_{\ell,\ell,i}^{(3,0)} + \tilde{\kappa}_{j,j,i}^{(3,0)} \cdot \tilde{\kappa}_{i,\ell,\ell}^{(3,0)} + \tilde{\kappa}_{j,j,i}^{(3,0)} \cdot \tilde{\kappa}_{\ell,i,\ell}^{(3,0)} + \\ \tilde{\kappa}_{j,j,i}^{(3,0)} \cdot \tilde{\kappa}_{\ell,\ell,i}^{(3,0)} \end{array} \right)}{72n} \quad (18)$$

which again, rather conveniently, can be combined with (13) and (17) since

$$\sum_{i,j,\ell=1}^k \dots = \sum_{i \neq j \neq \ell}^k \dots + \sum_{i=j \neq \ell}^k \dots + \sum_{i \neq j=\ell}^k \dots + \sum_{i=\ell \neq j}^k \dots + \sum_{i=\ell=j}^k \dots$$

resulting in

$$\sum_{i,j,\ell=1}^k (-''-) = \sum_{i \neq j \neq \ell}^k (-''-) + 3 \sum_{i=j}^k (-'''-) + 15 \sum_i^k \left( \kappa_{i,i,i}^{(3,0)} \right)^2 \quad (19)$$

where  $(-''-)$  refers to the big parentheses of (18) and  $(-'''-)$  to the big parentheses of (17). Due to the total symmetry of the cumulants, the full answer can be further simplified to

$$\sum_{i,j,\ell=1}^k \frac{\left( \tilde{\kappa}_{i,j,\ell}^{(3,0)} \right)^2}{12n} + \sum_{i,j,\ell=1}^k \frac{\tilde{\kappa}_{i,j,j}^{(3,0)} \cdot \tilde{\kappa}_{i,\ell,\ell}^{(3,0)}}{8n}$$

As before, increasing the upper limit of the summation from  $k$  to  $p$  changes nothing. And, since both terms are rotationally symmetric, the final version of the formula is

$$c \equiv \sum_{i,j,\ell=1}^p \frac{\left( \kappa_{i,j,\ell}^{(3,0)} \right)^2}{12n} + \sum_{i,j,\ell=1}^p \frac{\kappa_{i,j,j}^{(3,0)} \cdot \kappa_{i,\ell,\ell}^{(3,0)}}{8n} \quad (20)$$

using the original  $\mathbf{X}$  (not the rotated  $\mathbf{Z}$ ) cumulants.

## 4 Conclusion

We have thus found that the  $\frac{1}{n}$ -accurate CDF of  $T - d$  is given by

$$F(u) + \left[ a \cdot \left( -\frac{2u}{k} \right) + b \cdot \left( \frac{2u}{k} - \frac{2u^2}{k(k+2)} \right) + c \cdot \left( -\frac{2u}{k} + \frac{4u^2}{k(k+2)} - \frac{2u^3}{k(k+2)(k+4)} \right) \right] \cdot \frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2} \Gamma(\frac{k}{2})}$$

where  $a$ ,  $b$ ,  $c$  and  $d$  are defined in (9), (16), (20) and (10) respectively, and  $F(u)$  is the CDF of the  $\chi_k^2$  distribution defined in (7).

The corresponding PDF equals, by simple differentiation

$$\frac{u^{k/2-1} \exp(-\frac{u}{2})}{2^{k/2} \Gamma(\frac{k}{2})} \cdot \left[ 1 + a \cdot \left( \frac{u}{k} - 1 \right) + b \cdot \left( \frac{u^2}{k(k+2)} - \frac{2u}{k} + 1 \right) + \right. \\ \left. c \cdot \left( \frac{u^3}{k(k+2)(k+4)} - \frac{3u^2}{k(k+2)} + \frac{3u}{k} - 1 \right) \right]$$

when  $u > 0$  (zero otherwise). Note that the actual diagonalization of  $\mathbb{V}_0$  is needed only for establishing the  $a$  and  $d$  constants -  $b$  and  $c$  can be found directly from the 3<sup>rd</sup> and 4<sup>th</sup> cumulants of  $\mathbf{X}$ .

Examples of applying this procedure can be found in our two references; the second one incorrectly assumed that  $d = 0$  - with the help of this article, one can easily make the corresponding correction (it turns out that, in this particular case,  $a$  and  $d$  must have equal values).

And one final remark: it would prove rather difficult to extend this procedure to the full  $\frac{1}{n^2}$  accuracy; nevertheless, it usually proves beneficial to compute the value of each  $a$  and  $d$  to be  $\frac{1}{n^2}$  accurate - this tend to extend the applicability of the resulting approximation to such small sample sizes that further improvements do not appear (from a practical point of view) necessary.

## References

- [1] Duszak K and Vrbik J: "Improving Accuracy of Goodness-of-fit Test" *arXiv: 1410.6869* (Oct. 2014) 1-8
- [2] Vrbik J: "Improving Accuracy of Chi-squared Test of Independence" *Adv. Appl. Stat.* **39** #2 (2014) 81-94